A Pragmatic Approach to Enterprise-wide Data Lineage
Part 1: Fundamentals

A Global IDs Technical Whitepaper
October 15, 2019 (revised)

Arka Mukherjee, Ph.D., Kirk Kanzelberger, Ph.D., Bill Winkler, and Stefanos Damianakis, Ph.D.
info@globalids.com

This document is the first in a series of educational technical papers that we at Global IDs plan to release in the coming months, to share our experiences and lessons learned in enterprise data management with a wider public.

Introduction

Data lineage traces the *movement of information* within a firm's data ecosystem. In large enterprises, that ecosystem may comprise thousands of applications and systems of record, and tens of thousands of connections through which data flows between systems. In the absence of data lineage, there is no comprehensive view of the movement of information, but only the many "local" views of application owners who are intimately familiar with the details of their specific application's inbound and outbound feeds.

The mere aggregation of these "local" views — represented by a large collection of application-centric data flow diagrams, for example — cannot provide adequate support for key use cases such as the following:

- *Compliance*. The key to data compliance is explainability: the ability to trace data items back to their points of origin. Data dependency chains must be discovered and visualized, so that critical flow segments can be identified that affect any output that is subject to regulatory requirements.
- *Data quality.* Discovering and visualizing how data streams merge, split and are transformed by applications is key to identifying all the processes that affect data quality or are impacted by its degradation, as well as the critical flow points in the ecosystem where data quality controls should be placed.

- *Rationalization and refactoring.* Modernizing data architectures, eliminating inefficiency, and refactoring legacy applications are inherently complex processes that require rewiring data flows — breaking some existing flows and creating others. Data flow transparency is a key enabler for both planning and implementation stages of these processes.

- *Evolution of the data architecture.* Data environments undergo large scale change, driven by application development cycles, compliance initiatives, cloud migration, mergers, and so forth. Feasibility assessment, project planning, execution and oversight depend on detailed knowledge of data flows.

To date, the dominant approach to data lineage has relied on a *manual process*: conducting surveys of application owners, collating the resulting inventories of incoming and outgoing data feeds for every application, and then manually generating a comprehensive picture of data flow within the enterprise. The purely manual approach to data lineage is , in our opinion, inadequate for a number of reasons that we will explore shortly.

A *purely automated* approach to data lineage, however, is not merely inadequate but intractable, owing to the inherent complexity of the problem.

The approach to data lineage that we propose in this document is a *hybrid* approach that leverages human input, advanced algorithms, and automated methods of verifying and discovering data flows, that taken together render data lineage tractable and effective.

> **A note on terminology**. *Data lineage* in this whitepaper signifies a type of analysis that establishes *data provenance* in logical and actionable terms. Data lineage thus differs from the more limited conceptions of *data movement* and *data reconciliation* that are sometimes confused with it.
>
> In data lineage, the movement of information is traced through the data environment as a whole, from its point of origin to its terminal consumption point within the ecosystem. This tracing is framed in terms of business concepts in such a way that the provenance of individual data items of specific logical types can be understood and visualized. It thus goes beyond a more local analysis of ETL or message flows (data movement), as well as the mere comparison of source and target field values for purposes of integrity verification (data reconciliation).

The dimensions of the data lineage problem

All four use cases described above demand a *comprehensive view* of the movement of information. Without data lineage, efforts of these kinds — establishing data provenance, data quality control, streamlining and intelligent rewiring of data flows, the rationalization of complex environments, and the effective oversight of large scale change — are vulnerable to errors and inaccuracies stemming from a lack of knowledge of *how data actually flows* from end to end through the ecosystem. This, then, is the first dimension of the problem: *determining the actual existence and directionality of all of the data flows* between nodes or applications through the ecosystem.

However, for this comprehensive mapping of data flow to be truly actionable for business purposes, it must be *logically framed.* It is not enough to understand and visualize the flow of data in physical terms, or merely to know which columns of which tables are represented in the records flowing from system A to system B. What needs to be discovered and understood are *chains of data dependencies framed in terms of core business concepts* — the fundamental classes of entity (Customer, Product, User Account, Subscription, etc.) that comprise the business data for the particular enterprise. In other words, it is not merely the movement of data, but the movement of information pertaining to each specific core business concept that needs to be tracked and visualized. This is impossible without linking columns to concepts: to what semantic domain does a column belong, and to what business concept(s) does it pertain? *Accurate data classification* at the object and attribute levels is thus the second dimension of the problem.

This leads to the third dimension of the problem: however comprehensive high-level views of the movement of information might be, lack of depth renders these views effectively useless for detailed decision making and evidence-based explainability. The view of the flow of information provided by data lineage must be not only comprehensive, but sufficiently *granular*, reaching down to the object and attribute levels.

We summarize our position by saying that data lineage needs to be computed objectively, at the logical and physical levels, at the right level of granularity before it can be effectively used to solve the use cases outlined above.

Given these three dimensions of the problem, we can now explore and clarify the inadequacies of purely manual approaches to data lineage, the intractability of a purely automated approach, and the path to a solution via a hybrid approach that leverages the strengths of both human and machine intelligence.

Inadequacies of a purely manual process for data lineage

Purely manual approaches to data lineage are affected by several grave inadequacies:

- *Fallibility*. The source data for a purely manual approach principally consists in the collated responses of application owners to requests for information about the feeds into and out of their applications. While many application owners are intimately familiar with the details of inbound and outbound flow, survey responses will still contain inaccuracies, inadvertent omissions, unwitting reliance on incomplete or outdated information, and guesswork.

- *Data classification is not scalable in human terms.* Humans can define the taxonomy of business concepts and indicate which concepts may be implicated in the flow of data into and out of a given application. But the linchpin of the logical depiction of data flow is *data classification* that links individual attributes to concepts. Data classification must be automated, since the number of database columns over all applications will number in the millions, or in the tens of millions for a large enterprise. See Figure 1 below.

- *Restriction to high-level information*. The collation of responses from application owners may enable a high-level depiction of data flow, but does not provide the object- and attribute-level granularity necessary for rigorous tracking of data in terms of business concepts.

- *Lack of objective verification.* Lastly, human claims about the origin and flow of data are just that: human claims. Claims that are backed up by evidence computed from the data itself are far stronger, far more likely to satisfy regulators, and far more secure as a basis for decision-making than assurances on paper that lack objective verification.

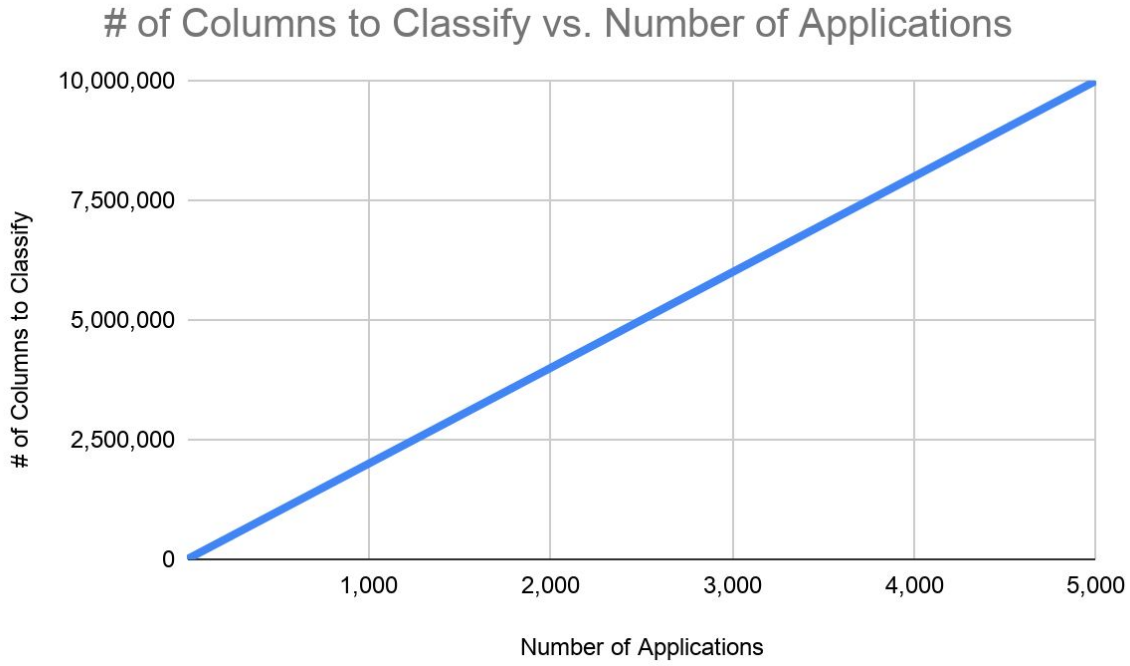

Figure 1. How the Number of Columns to Classify Scales with the Number of Applications

This discussion of the inadequacies of a purely manual process invites the question: can data lineage simply be automated?

The impracticability of a purely automated approach to data lineage

As attractive as it may sound, a purely automated approach to data lineage turns out to be impractical. Human intelligence excels at high-level description and analysis. Without human inputs, a purely automated approach to data lineage must proceed by a *method of exhaustion*: attempting to discover data flow and dependencies by checking for the presence of every possible connection between applications in the ecosystem.

Aside from the kind of universal access that would be required to carry out this method of exhaustion, the complexity of an exhaustive search for flow segments between thousands of applications becomes computationally prohibitive. To understand why this is the case, consider this simple diagram comprising a very small number of systems:

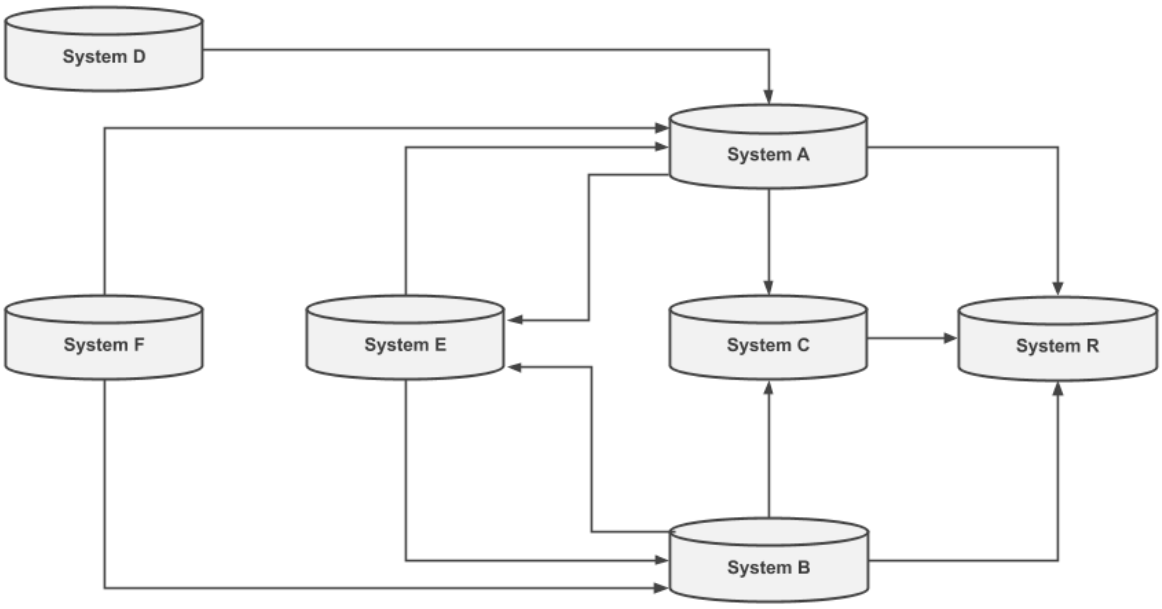

Figure 2. Example data flow diagram for an environment comprising seven systems.

Figure 2 depicts a data environment comprising a total of seven systems and twelve flow segments. Note that each flow segment is directional: for example, the segment connecting System D and system A is a data feed outbound from D and inbound to A. System E and System B are connected by two flow segments: one flowing from E to B and another flowing from B to E.

A purely automated approach to discovering data flow in an ecosystem comprising seven applications would have to check for the presence of a connection in either direction for a total of (7 × 6) / 2 = 21 distinct pairs of systems. In other words, it would have to check for 21 × 2 = 42 possible flow segments in order to discover the 12 actual flow segments.

Now consider how the complexity grows for a purely automated approach when the number of systems is in the thousands. Applying a similar calculation as in the above example, a data ecosystem with 5,000 applications has nearly 25,000,000 possible flow segments. In short, the flow segment complexity grows in proportion to the *square* of the number of systems involved. The scaling problem is depicted below in Figure 3.

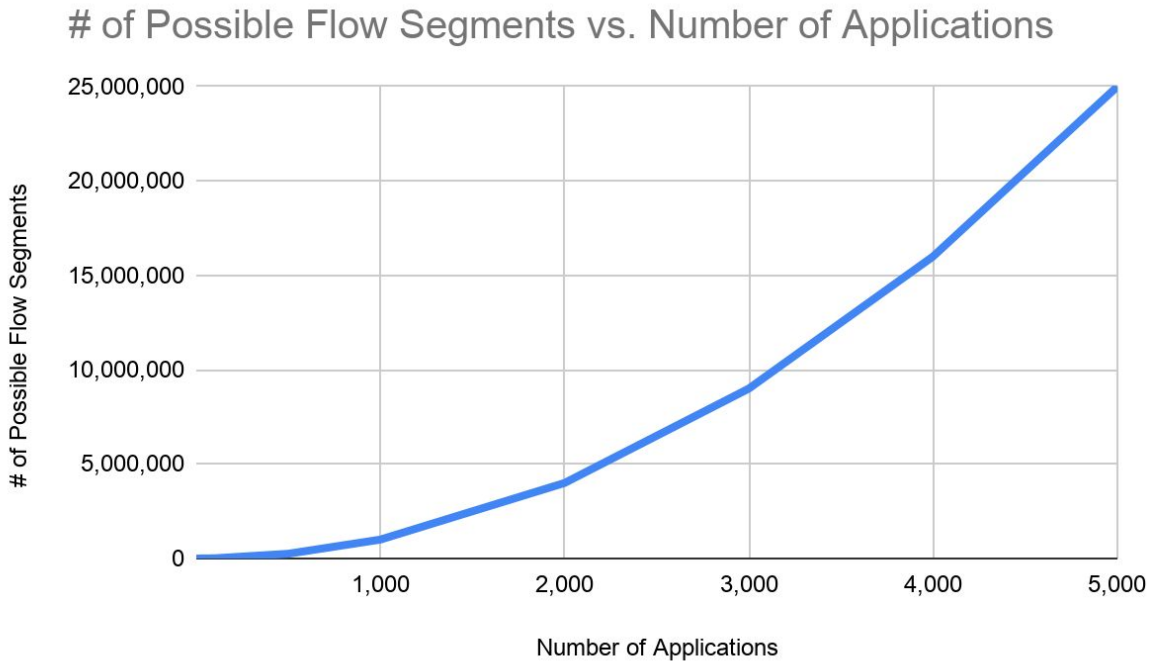

Figure 3. How Possible Flow Segments Scale with the Number of Applications

In the next section, we will outline some of the methods used to check for the existence of flow between systems, which involve sampling procedures applied to the rows and columns — which may number in the thousands — of database tables across pairs of applications. The time consumed by these procedures is significant. When we consider how the complexity scales upward with the number of systems, we can see the impracticability of a purely automated approach to data lineage.

Making data lineage tractable: A hybrid of human and machine intelligence

The key to a better strategy for data lineage lies in leveraging the strengths of both human and machine intelligence in a hybrid approach.

The human contribution

Human intelligence fulfills two important functions in this hybrid approach.

1. *Narrowing the flow-search problem*. As we have said, human intelligence excels in high-level description and analysis — "the big picture." Despite its fallibility and lack of granularity, the great advantage of human intelligence here lies in narrowing the universe of possible flows between systems, ameliorating the scaling issue discussed above.

While the number of *possible* flows is proportional to the square of the number of systems, there *actually* exist a manageable number of verifiable flows out of each system. Surveys and interviews of application owners serve, in the first place, to piece together the general data flow architecture, enabling the exclusion of a large portion of the universe of possibilities.

When pieced together, the high-level human descriptions of data flow serve in this hybrid approach as a set of *flow hypotheses* that are then subject to machine validation.

2. *Logical description of enterprise data in terms of business concepts*. The second important function of human intelligence in this context is to define the set of core business concepts and indicate their presence across the data landscape.

Why is this important? Business concepts are the keys for interpreting the *logical meaning* of flow segments and their business *context* (for example, how a particular segment may qualified by geography, or line-of-business). They enable an understanding of how logical meaning and context change as applications merge, split, and transform data from end to end on its journey through the ecosystem. Business concepts determine the set of objects and attribute types to be modeled by the software that performs automated data classification.

The machine contribution

Given human data flow hypotheses and a logical description of enterprise business concepts, the machine now makes its contribution to the hybrid approach. This has a number of aspects.

1. *Automated, scalable data classification (linking columns to business concepts)*. While a human can look at sets of columns and spot relationships to logical concepts, manual classification of millions or tens of millions of columns is clearly impracticable. The multi-class classification decisions can, however, be modeled using machine learning algorithms. As we have pointed out, the logical classification of data according to business concepts is the linchpin of a meaningful and actionable, as well as comprehensive, view of the movement of information.

2. *Verification of data flow segments.* Human inputs have reduced the universe of flow possibilities to a manageable set. The first important function of the machine is to verify that these hypothetical flow segments between applications and systems actually exist (along with the hypothesized directionality of the flow), and that the semantics of the machine-classified

data flowing in these segments matches the human description. *Consistent record sampling* in conjunction with automated data classification can detect objects in tables that fall under particular business concepts, thus verifying flow points for that concept through the ecosystem.

The data flow itself can be verified by a number of machine techniques.

- In *timestamp analysis*, records are selected in source and destination that have columns mapping to a specific semantic domain or business concept (as determined by data classification), and that contain timestamp information, such as a "Last Update Date/Time" field. The software then looks for intersections of these records across the two systems and sequences them according to the timestamp information. Sequencing patterns showing similar time intervals indicate both the existence and direction of flow between source and destination.

  Timestamp information, however, is often missing and can be unreliable (sometimes the "timestamps" are simply replicated from source to destination systems!). In such cases, stroboscopic analysis can be used to verify data flow.

- *Stroboscopic analysis* uses *differential record sampling* at regular intervals to observe flow and determine directionality. The software performs consistent record sampling of both source and destination at roughly the same time, using similar column selection criteria to those in timestamp analysis. It then repeats this sampling at a chosen "strobing" interval for some number of iterations. Similar to a stroboscope, this kind of analysis takes snapshots of data in motion to discover patterns of record updates or appearances of new records in the destination system that match data in the source system.

  Figure 4 below depicts differential sampling in stroboscopic analysis. Values appearing in the first database on the left consistently appear in the second database on the right the next day. Regular patterns of inserts or updates across two or more columns are evidence of a data flow segment. These segments can then be stitched together to form lineage flow diagrams.

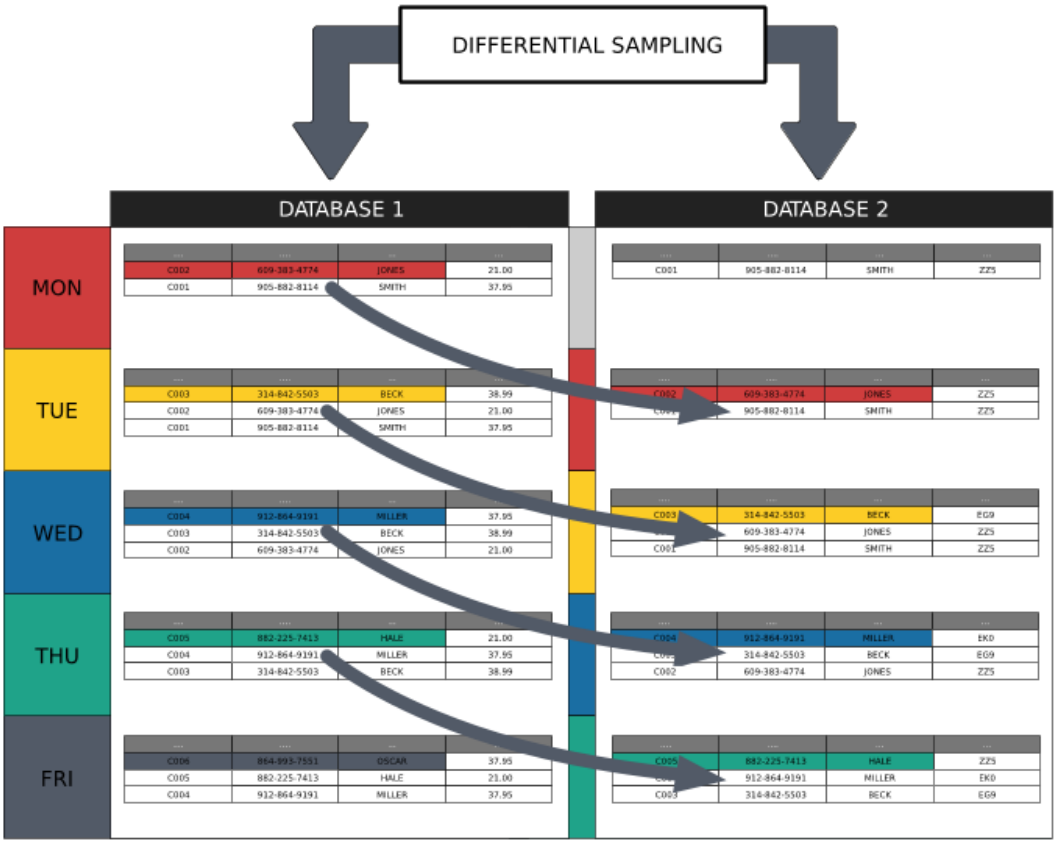

Figure 4. DIfferential Record Sampling.

3. Discovery of data flow segments. The same techniques that are used to verify human hypotheses concerning data flow can be used to discover undocumented data flow segments. Rather than pursuing a method of exhaustion as a wholly automated approach would attempt to do, high-level human hypotheses concerning data flow serve here as a guide for additional machine-generated hypotheses that may discover unreported flow segments or provide evidence correcting errors in human data flow hypotheses.

Additional techniques for automated flow discovery and verification include scanning of stored procedures in databases and ETL code. It is not necessary to parse stored procedures or ETL code fully to discover references to data and CRUD operations on other systems, which provide evidence of flow segments that can be corroborated using the sampling techniques already mentioned.

4. *Granular visualization of data lineage.* Machines excel at drill-down, collation, and compilation of large volumes of detailed information beyond the capacity of humans. Regulatory reporting, data governance, and ecosystem development decisions all benefit from the ability to visualize data flow across the ecosystem logically (framed by business concepts) and at increasing levels of granularity, down to the object and attribute levels. In enabling this kind of data transparency, machines are conscripted do the hard work, guided by the conceptual data models and high-level descriptions of flow provided by humans.

Summary

It is our belief that a hybrid approach to data lineage that leverages the strengths of both human intelligence, data classification, and automated flow discovery is best equipped to deliver a comprehensive, meaningful, and actionable view of the movement of data in large ecosystems.

Understanding data flows comprehensively and logically in terms of business concepts is necessary for any meaningful discussion of data provenance, for identifying and tracking data subject to regulation wherever it occurs in the ecosystem, for understanding the implications of rewiring data flows, for assessing efficiencies and inefficiencies of the enterprise-wide data architecture, and for accurately evaluating proposals for rationalizing and streamlining the data environment. Human intelligence defines the semantics, and data classification uses machine learning to automate the linkage of millions of columns of data to business concepts.

Human intelligence likewise narrows the universe of possible flow segments in such a way that automated flow verification and discovery become practicable. The object- and attribute-level granularity available in this hybrid approach is indispensable for any rigorous demonstrations of data provenance and informed decision-making in the governance and evolution of the data architecture in large enterprises.

In the second part of this whitepaper, we discuss the implementation of this hybrid approach.